2014 European Control Conference (ECC)
June 24-27, 2014. Strasbourg, France

Bayesian and nonparametric methods for system identification and model selection

A. Chiuso[†] and G. Pillonetto[†]

Abstract— **System Identification has been developed, by and large, following the classical parametric approach. In this tutorial we shall discuss how Bayesian statistics and regularization theory can be employed to tackle the system identification problem from a nonparametric (or semi-parametric) point of view. The present paper provides an introduction to the use of Bayesian techniques for smoothness and sparseness, which turn out to be flexible means to face the bias/variance dilemma and to perform model selection.**

Index Terms— **Nonparametric methods, Sparsity, kernel Methods, Sparse Bayesian Learning, Optimization**

I. INTRODUCTION

System Identification is concerned with automatic model building from measured data. Under this unifying umbrella, this field spans a rather broad spectrum of topics, considering different model classes (linear, hybrid, non-linear, continuous and discrete time) as well as a variety of methodologies and algorithms, bringing together in a nontrivial way concepts from classical statistics, machine learning and dynamical systems.

Even though considerable effort has been devoted to specific areas, such as parametric methods for linear system identification which are by now well developed (see [19], [26]), it is fair to say that modeling still is, by far, the most time consuming and costly step in Advanced Process Control applications. As such, the demand for fast and reliable automated procedures for system identification makes this exciting field still a very active and lively one.

Suffices here to recall that, following this classic parametric Maximum Likelihood (ML)/Prediction Error (PE) framework, the candidate models are described using a finite number of parameters $\theta \in \mathbb{R}^n$. After the model classes have been specified, the following two steps have to be undertaken:

(i) estimate the model complexity $\hat{n}$
(ii) find the estimator $\hat{\theta} \in \mathbb{R}^{\hat{n}}$ minimizing a cost function $J(\theta)$, e.g. the prediction error or (minus) the log-likelihood.

Both of these steps are critical, yet for different reasons: step (ii) boils down to an optimization problem which, in general, is non-convex and as such it is very hard to guarantee that a global minimum is achieved. The regularization techniques discussed in this paper sometimes allow to reformulate the identification problem as a convex program, thus solving the issue of local minima.

In addition fixing the system complexity equal to the "true" one[1] is a rather unrealistic assumption and in practice the complexity n has to be estimated as per step (i); this is classically performed using model order selection criteria such as AIC, BIC, MDL or cross validation techniques [19], [26]. This has non-trivial implications, chiefly the facts that classical order selection criteria are based on asymptotic arguments and that the statistical properties of estimators $\hat{\theta}$ after model selection, called Post Model Selection Estimators (PMSE), are in general difficult to study [18] and may lead to undesirable behavior. Experimental evidence shows that this is not only a theoretical problem but also a practical one [21], [6]. On top of this statistical aspect there is also a computational one. In fact the model selection step, which includes as special cases also variable selection and structure selection, may lead to computationally intractable combinatorial problems. Two simple examples which reveal the combinatorial explosion of candidate models are the following: (a) *Variable Selection*: consider a high dimensional time series (MIMO) where not all inputs/outputs are relevant and one would like to select k out of m available input signals where k is not known and needs to be inferred from data, see e.g. [3], [8]); (b) *Structure selection*: consider all autoregressive models of maximal lag p with only $p_0 < p$ non-zero coefficients and one would like to estimate how many (p_0) and which coefficients are non-zero. Given that enumeration of all possible models is essentially impossible due the combinatorial explosion of candidates, selection could be performed using greedy approaches from multivariate statistics, such as stepwise methods [16].

The system identification community, inspired by work in statistics [27], [20], machine learning [25], [28], [2] and signal processing [12], [32], has recently developed and adapted methods based on regularization to jointly perform model selection and estimation in a computationally efficient and statistically robust manner. Different regularization strategies have been employed which can be classified in two main classes: regularization induced by so-called smoothness priors (aka Tikhonov regularization, see [17], [11] for early references in the field of dynamical systems) and regularization for selection. This latter is usually achieved by

[†] A. Chiuso and G. Pillonetto are with the Dept. of Information Engineering, University of Padova, {chiuso,giapi}@dei.unipd.it

This work has been partially supported by the FIRB project "Learning meets time" (RBFR12M3AC), European Community's Seventh Framework Programme [FP7/2007-2013] underagreement n. 257462 HYCON2 Network of excellence.

[1] In practice there is never a "true" model, certainly not in the model class considered. The problem of statistical modeling is first of all an approximation problem; one seeks for an approximate description of "reality" which is at the same time simple enough to be learned with the available data and also accurate enough for the purpose at hand.

convex relaxation of the ℓ_0 quasi-norm (such as ℓ_1 norm and variations thereof such as sum-of-norms, nuclear norm etc.) or other non-convex sparsity inducing penalties which can be conveniently derived in a Bayesian framework, aka Sparse Bayesian Learning (SBL), [20], [28], [32].

The purpose of this paper is to guide the reader through the most interesting and promising results on this topic as well as areas of active research; of course this subjective view only reflects the authors' opinion and of course different authors could have offered a different perspective. We also refer the reader to [24] for a recent survey.

II. Bayesian approach to System Identification

Let us consider, for the sake of exposition, only Output Error models. The extension to more general model classes can be found in [21], [6], [8] and references therein. Therefore we consider models of the form

$$y(t) = [h * u](t) + e(t) \quad y(t) \in \mathbb{R}^p \tag{1}$$

where $h(t) \in \mathbb{R}^{p\times m}$ is the impulse response of the system, $u(t) \in \mathbb{R}^m$ is the measurable input and $e(t)$ is a zero mean Gaussian white noise process uncorrelated from $u(t)$; we consider both $t \in \mathbb{R}$ for continuous time systems and $t \in \mathbb{Z}$ for discrete time systems. The symbol $[h * u](t)$ denotes convolution which is a linear operator mapping the impulse response h onto outputs y:

$$[h * u](t) = \mathscr{L}_u[h] : h(t) \to y(t)$$

The linear operator $\mathscr{L}_u$ is completely specified by the input process $u(t)$. The problem of system identification is to estimate the (matrix valued) function $h(t)$ starting from a finite set of input output data points $\{u(t), y(t)\}_{t\in\mathscr{T}}$. Here $\mathscr{T}$ denotes a discrete set of time instants where measurements are available. Without loss of generality we assume $\mathscr{T} := \{0, 1, 2, ..., T-1\}$.

In the Bayesian (or regularization) approach to system identification one postulates a probability description of the unknown impulse response[2]. Let us say that

$$h \sim p_\eta(h) \tag{2}$$

where the prior $p_\eta(h)$ may depend upon some unknown parameters (hyperparameters hereafter) which need to be estimated from data.

One typical and convenient choice is to postulate that $h(t)$ is a zero mean Gaussian process [25], independent of the noise $e(t)$ with covariance function $K_\eta(t,s) := \mathbb{E}h(t)h(s)$, which is sometimes called kernel in the Machine Learning community. In the rest of the paper we shall interchangeably encode the prior model on $h(t)$ either providing a description of the Gaussian process in terms of (stochastic) differential equations, or via its covariance function/kernel.

[2]This may be a delicate point from the probabilistic point of view since $h(t)$ is an infinite dimensional object. We shall skip the technical details here.

Let us denote with Y the vector with components $y(t)$, $t \in \mathscr{T}$. Since h and e are Gaussian and independent, and $\mathscr{L}_u[h]$ is linear, then Y and $h(t)$ are jointly Gaussian so that[3]

$$p_\eta(h|Y) := \frac{p(Y|h)p_\eta(h)}{p_\eta(Y)} \quad p_\eta(Y) = \int p(Y|h)p_\eta(h)\,dh \tag{3}$$

is Gaussian. Hence, the posterior estimate

$$\hat{h} := \mathbb{E}_\eta[h|Y] \tag{4}$$

can be computed in closed from for any fixed η.

One popular approach to estimating η is to maximize the so-called marginal likelihood $p_\eta(Y)$, i.e. the likelihood of the hyper parameters η once the unmeasurable quantities (h) have been integrated out. Let us denote with

$$\hat{\eta} := \arg\max_\eta \; p_\eta(Y) \tag{5}$$

the marginal likelihood estimator of η. Then replacing η in (4) with its estimate $\hat{\eta}$ (5) we obtain the so-called empirical Bayes estimator of h:

$$\hat{h} := \mathbb{E}_{\hat{\eta}}[h|Y] \tag{6}$$

In the remaining part of the paper we shall discuss the design of a prior $p_\eta(h)$. For Gaussian priors this will be translated into the problem of designing the kernel $K_\eta(t,s)$. For convenience of notation we shall first assume that the system is continuous time, i.e. $t \in \mathbb{R}$. The kernel for the discrete time problem can be found sampling the continuous time version.

It is useful to observe that another commonly used estimator of h is the so-called MAP estimator, i.e.

$$\hat{h}_{MAP} := \arg\max_h p_{\hat{\eta}}(h|Y) \tag{7}$$

which clearly coincides with (6) if the posterior density $p_{\hat{\eta}}(h|Y)$ is unimodal and symmetric around its mean (which holds, e.g., if h is, conditionally on Y, a Gaussian random process). A simple application of the Bayes rule shows that $\hat{h}_{MAP}$ can also be written as

$$\hat{h}_{MAP} \quad = \quad \arg\min_h J_F(h) + J_R(h;\hat{\eta}) \tag{8}$$

where

$$J_F(h) := -\log(p(Y|h)) \quad J_R(h;\eta) := -\log(p_\eta(h)) \tag{9}$$

are, respectively, the "Fit" and "Regularization" terms. The first measures how well the model h describes the data Y while the second penalizes certain "unlikely" systems h. Equation (8) can be seen as a way to deal with the *bias-variance tradeoff*. The regularization term $J_R(h;\eta)$ may depend upon some regularization parameters η (called also hyper parameters) which need to be tuned using measured data. The Bayesian interpretation (see (9), (3)) offers one possible route to estimating the hyper parameter vector η

[3]Note that the conditional $p(Y|h)$ is just the noise density. For simplicity here we assume the noise variance σ^2 to be known and fixed. Of course one can also include σ^2 in the hyper parameter vector η and estimate it using (5).

via marginal likelihood optimization (5). From now on we shall interchangeably use $J_F(h)$ and $J_R(h;\eta)$ or $p(Y|h)$ and $p_\eta(h)$ which will be always linked as in (9). We now discuss different forms of regularization $J_R(h;\eta)$ which have been studied in the literature.

The prior $p_\eta(h)$ (or equivalently the regularization term $J_R(h;\eta)$) can be roughly classified in *regularization for smoothness*, which attempts to control complexity in a smooth fashion and *regularization for sparseness* which, on top of estimation, also aims at selecting among a finite (yet possibly very large) number of candidate model classes.

We shall first discuss *regularization for smoothness* (see Section III), introducing a family of kernels $K_\eta(t,s)$ known as stable spline kernels and then *regularization for sparseness*, see Section IV.

III. Priors for smoothness and Tikhonov regularization

Let us consider the OE model (1) and further assume that $h(k)=0$, $\forall k>T$, so that $H(q):=\sum_{k=1}^{T} h(k)z^{-k}$. Let h be the vector containing all the unknown coefficients of the impulse response $\{h(k)\}_{k=1,\dots,T}$ and define $\hat{y}_{t|t-1}(h):=\sum_{k=1}^{T} h(k)u(t-k)$. Define also $y\in\mathbb{R}^N$ be the vector of output observations, Φ the regressor matrix with past input samples and e the vector with innovations (zero mean, variance $\sigma^2 I$). With this notation the convolution input-output equation (1) takes the form

$$Y=\Phi h+e$$

The linear least squares estimator

$$\begin{array}{rcl}\hat{h}_{LS} & := & \arg\min_h \quad J_F(h)\\ J_F(h) & := & \frac{1}{N}\sum_{t=1}^{N}\|y_t-\hat{y}_{t|t-1}(h)\|^2\end{array} \tag{10}$$

is ill-posed unless the number of data N is larger (and in fact much larger) that the number of parameters T. From the statistical point of view the estimator (10) would result, for large T in small bias and large variance. The purpose of regularization is to render the inverse problem of finding h from the data $\{y_t\}_{t=1,\dots,N}$ well posed, thus better trading bias versus variance. The simplest form of regularization is indeed the so called ridge-regression or its weighted version (aka generalized Tikhonov regularization), where the 2-norm of h is weighted w.r.t. a positive semidefinite matrix K_η,

$$\begin{array}{rcl}\hat{h}(\eta) & := & \arg\min_h \quad J_F(h)+J_R(h;\eta)\\ J_R(h;\eta) & := & h^\top K_\eta^{-1} h\end{array} \tag{11}$$

For system identification problems the matrix K_η, aka kernel, will have to capture specific properties of impulse responses (exponential decay, BIBO stability, smoothness etc. [23], [10]). Early references include [11], [17], while more recent work can be found in [21], [6] where several choices of kernels are discussed, see Section III-A for more details.

For this choice of J_F and J_R, and having fixed η, the estimator $\hat{h}(\eta)$ is the solution of a quadratic problem and can be written in closed form (aka Ridge Regression):

$$\hat{h}(\eta)=K_\eta\Phi^\top\left(\Phi K_\eta\Phi^\top+\sigma^2 I\right)^{-1}Y \tag{12}$$

Two common strategies adopted to estimate the parameters η are Cross Validation [19] and marginal Likelihood maximization. This latter approach is based on the Bayesian interpretation given in equations (3) from which one can compute the so called "Empirical Bayes" estimator $\hat{h}:=\hat{h}(\hat{\eta})$ (6) of h plugging the maximum marginal Likelihood (5) estimator $\hat{\eta}$ in (12). The main strength of the marginal likelihood is that, by integrating the joint posterior over the unknown h, it automatically accounts for the residual uncertainty in h for fixed η. When J_F and J_R are quadratic as in (10), (11), which according to (9) corresponds to assuming that e and h are independent and Gaussian, the marginal likelihood in (5) can be computed in closed form so that

$$\begin{array}{rcl}\hat{\eta} & := & \arg\min_\eta \quad \log(\det(\Sigma(\eta))+Y^\top\Sigma^{-1}(\eta)Y\\ \Sigma(\eta) & := & \Phi K_\eta\Phi^\top+\sigma^2 I\end{array} \tag{13}$$

It is here interesting to observe that $\hat{\eta}$ which solves (5), under certain conditions, leads to $K_{\hat{\eta}}=0$ (see example 1 in Section IV), so that the estimator of h in (12) satisfies $\hat{h}(\hat{\eta})=0$. This simple observation is the basis of so-called *Sparse Bayesian Learning* (SBL); we shall return on this issue in the next section when discussing regularization for sparsity and selection.

Unfortunately the optimization problem (5) (or (13)) is not convex and thus subjected to the issue of local minima. However, both experimental evidence as well as some theoretical results support the use of marginal likelihood maximization for estimating regularization parameters, see e.g. [25], [1].

A. Stable Kernels

One of the major breakthroughs in [23] has been to introduce a class of prior description for impulse responses which encoded structural properties of dynamic systems such as BIBO stability. Even though several kernels have been later introduced in the literature such as TC/DC [6] and multiple versions [4], we shall here consider only one particular instance named stable spline kernel. Let us first introduce the stable spline kernel in continuous time as follows. Let $w(\tau)$ be the normalized Wiener process[4] which is defined as a Gaussian zero mean process which satisfies:

$$\begin{array}{l}w(0)=0\\ w(\tau)-w(s)\sim\mathcal{N}(0,\tau-s)\quad \tau\geq s\geq 0\\ w(\tau_2)-w(\tau_1)\perp w(s_2)-w(s_1)\qquad \tau_2\geq\tau_1\geq s_2\geq s_1\geq 0\end{array}$$

The covariance of the normalized Wiener process is

$$r(\tau,s)=\mathbb{E}w(\tau)w(s)=\min(\tau,s)\quad \tau\geq 0,\ s\geq 0$$

While widely used in Machine Learning [25] and Statistics [29] as a reasonable prior for function estimation, it is not suitable for describing the impulse response $h(t)$, $t\in\mathbb{R}$ of a BIBO stable linear system which should be an absolutely integrable function. To this purpose [23] have introduced an

[4]The Wiener process can also be informally defined as integrated (continuous time) "white" Gaussian noise.

exponential time change $\tau = e^{-\beta t}$ which maps $t \in \mathbb{R}^+ := [0,+\infty)$ onto $\tau \in (0,1]$ and defined

$$w_\beta(t) := w(e^{-\beta t}) \quad t \in \mathbb{R}^+ \quad \in (0,+\infty) \tag{14}$$

whose covariance is given by

$$r_\beta(t,s) = \mathbb{E} w_\beta(t) w_\beta(s) = \min(e^{-\beta t}, e^{-\beta s}) = e^{-\beta \max(t,s)} \tag{15}$$

It has been shown in [23] that realizations from (14) are almost surely absolutely integrable. Indeed, extending ideas from [9], it can be shown that (14) can be thought of as the "Maximum Entropy" prior model under a finite variance constraint of its first order derivative, see [22].

Remark 1: Note that the exponential decay of $h(t)$ guarantees that, to any practical purpose, it can be considered zero for $t > T$ for a suitably large T. This allows to approximate the OE model (1) with a "long" *Finite Impulse Response* (FIR) model. In discrete time $t \in \mathbb{Z}$ the impulse response $h(k)$, $k = 1,..,T$ is now modeled as zero mean Gaussian vector with covariance $cov(h(t),h(s)) := K_\eta(t,s)$.

This prior has been further enriched in [21] by adding a parametric component which has some advantages when describing fast oscillating systems. This is indeed the basic building bock of other multiple kernels (see e.g. [7]) which, for reasons of space, cannot be surveyed here. We shall now take a small detour on the structure of the process (14). This, hopefully, will allow us to fully understand the prior, as well as will guide us through possible modifications which make it more suitable to describing linear systems.

B. The Wiener process and the Brownian Bridge

It is a a result by Wiener (see e.g. [31]) that the Wiener process $w(\tau)$, $\tau \in [0,1]$, admits the (almost sure) random Fourier series expansion:

$$\begin{aligned} w(\tau) &= \psi_0 \tau + \sqrt{2}\sum_{k=1}^{\infty} \psi_k \frac{\sin(\pi k \tau)}{\pi k \tau} \quad \psi_k \sim \mathcal{N}(0,1) \ \ i.i.d. \\ &= \psi_0 \tau + b(\tau) \end{aligned} \tag{16}$$

where the last equation defines $b(\tau)$. It is a simple check that $w(1) = \psi_0$ so that equation (16) induces a decomposition of $w(\tau)$ into a linear term $\psi_0 t$, which is the conditional mean of $w(t)$ given its value at one

$$\mathbb{E}[w(\tau)|w(1) = \psi_0] = \psi_0 \tau,$$

and a term $b(\tau)$ which vanishes at the extremes of the interval, i.e. $b(0) = b(1) = 0$. The process $b(\tau)$ is called, in the stochastic process literature, *Brownian Bridge*, which models the "erratic behavior" of $w(\tau)$ around its conditional mean $\psi_0 \tau$ given its value $\psi_0 := w(\tau)_{|\tau=1}$.

C. Linear systems and the β-exponential Brownian Bridge

We now take a closer look at (14) ant its relation with $b(\tau)$. Consider now the random Fourier series (16) and introduce the exponential time change $\tau := e^{-\beta t}$. It follows that

$$\begin{aligned} w_\beta(t) &:= w(e^{-\beta t}) \quad t \in \mathbb{R}^+ \quad \beta \in (0,+\infty) \\ &= \psi_0 e^{-\beta t} + b(e^{-\beta t}) \\ &= \psi_0 e^{-\beta t} + b_\beta(t) \end{aligned} \tag{17}$$

where the last equation defines $b_\beta(t)$ which we call the β-*exponential Brownian Bridge*. The process $b_\beta(t)$ inherits the property of $b(\tau)$ which vanishes at the extremes of $[0,1]$, so that $b_\beta(t)$ vanishes at the extremes of $[0,+\infty)$, i.e.

$$b_\beta(0) = \lim_{t\to\infty} b_\beta(t) = 0 \quad a.s.$$

This implies that ψ_0 represents the value at $t = 0$ of $w_\beta(t)$.

The decomposition in equation (17) provides a nice interpretation in terms of linear system theory as follows. Taking $w_\beta(t)$ as a prior model of an impulse response $h(t)$ amounts to assuming that $h(t)$ can be modeled as the sum of an exponential function $\psi_0 e^{-\beta t}$, which also encodes the value at zero of the impulse response through ψ_0, and the $\beta-$exponential Brownian Bridge which describes fluctuations around $\psi_0 e^{-\beta t}$ (conditionally on $\psi_0 := w_\beta(0) = h(0)$).

The "variation" $b_\beta(t)$ around the conditional mean $\psi_0 e^{-\beta t}$ allowed for by the prior (14) is determined by the random variables ψ_k, $k = 1,...,+\infty$; recall from (16) that $Var\{\psi_k\} = 1$, $\forall k \geq 0$. This indicates that this prior rigidly links the variance of ψ_0 (which is needed to encode the value at zero of the impulse response) with the variance of the stochastic component $b_\beta(t)$ describing the variation around the conditional mean. This suggests a first extension of the stable spline model (14) allowing for different (nonnegative) weightings of the conditional mean and the Browinian Bridge:

$$w_{\beta,\gamma}(t) := \sqrt{\gamma_e}\psi_0 e^{-\beta t} + \sqrt{\gamma_b} b_\beta(t) \quad \gamma_e, \gamma_b \in [0,+\infty) \tag{18}$$

The covariance of (18) is thus given by:

$$K_{\beta,\gamma}(t,s) = \gamma_e e^{-\beta(t+s)} + \gamma_b \mathbb{E} b_\beta(t) b_\beta(s) \tag{19}$$

which is the sum of a rank-1 kernel $\gamma_e e^{-\beta(t+s)}$ (see also [5]) and a full rank kernel $\gamma_b \mathbb{E} b_\beta(t) b_\beta(s)$.

Note also that in (18) the "exponential" component $\psi_0 e^{-\beta t}$ is only able to describe one mode of a linear system. Thus impulse response of an $n-th$ order linear system, which is the superposition of n (possibly sinusoidally modulated) exponentials, needs to be described by the β^*-exponential Brownian Bridge component as variation around a sort of "average" exponential $\hat{\psi}_0 e^{-\beta^* t}$, for some choice of β^* which need not correspond to any of the system poles. This suggests that the single kernel (19) may be too rigid and modification containing multiple exponentials and/or sinusoidally modulated exponentials may be advantageous; such extensions are discussed in [7].

IV. Regularization for sparsity: variable selection and order estimation

The main purpose of regularization for sparseness is to provide estimators $\hat{h}$ in which subsets or functions of the estimated parameters are equal to zero.

Consider now the Multi Input Multi Output OE model (1) and denote with $y_j(t)$ the j-th component of $y_t \in \mathbb{R}^p$; let also $h \in \mathbb{R}^{T(m+p)}$ be the vector containing all the impulse response coefficients $h_{ij}(k)$, $j = 1,..,p$, $i = 1,..,m$ and $k =$

$1,..,T$. Simple examples of sparsity one may be interested in are (see also [8], [15]):

(i) single elements of the parameter vector h, which corresponds to eliminating specific lags of some variables;
(ii) groups of parameters such as the impulse response from i-th input to the j-th output $h_{ij}(k)$, $k = 1,..,T$, thereby eliminating the i-th input from the model for the j-th output
(iii) the singular values of the Hankel matrix $\mathscr{H}(h)$ formed with the impulse response coefficients $h(k)$; in fact the rank of the Hankel matrix equals the order (i.e. the McMillan degree) of the system[5].

To this purpose one would like to penalize the number of non-zero terms (let them be entries of h, groups, singular values etc). This is measured by the ℓ_0 quasi-norm or its variations (group ℓ_0 and ℓ_0 quasi-norm of the Hankel singular values, i.e. the rank of the Hankel matrix). Unfortunately if J_R is a function of the ℓ_0 quasi-norm the resulting optimization problem is computationally intractable; as such one usually resorts to relaxations. Three common ones are described below.

One possibility is to resort to greedy algorithms such as Orthogonal Matching Pursuit; generically it is not possible to guarantee convergence to a global minimum point.

A very popular alternative is to replace the ℓ_0 quasi-norm by its *convex envelope*, i.e. the ℓ_1 norm, leading to algorithms known in statistics as LASSO [27] or its group version Group LASSO [33]:

$$J_R(h;\eta) = \eta\|h\|_1 \tag{20}$$

Similarly the convex relaxation of the rank (i.e. the ℓ_0 quasi-norm of the singular values) is the so-called nuclear norm (aka Ky-Fan n-norm or trace norm), which is the sum of the singular values[6] $\|A\|_* := trace\{\sqrt{A^\top A}\}$ where $\sqrt{\cdot}$ denotes the matrix square root which is well defined for positive semidefinite matrices. In order to control the order (McMillan degree) of a linear system, which is equal to the rank of the Hankel matrix $\mathscr{H}(h)$ built with the impulse response described by the parameter h, it is then possible to use the regularization term

$$J_R(h;\eta) = \eta\|\mathscr{H}(h)\|_* \tag{21}$$

thus leading to convex optimization problems [14]. Both (21) and (20) induce sparse or nearly sparse solutions (in terms of elements or groups of h (20) or in terms of Hankel singular values (21)), making them attractive for selection. Yet, as well documented in the statistics literature, both (21) and (20) do not provide a satisfactory tradeoff between sparsity and shrinking, which is controlled by the regularization parameter η. As η varies one obtains the so-called *regularization path*. Increasing η the solution gets sparser but, unfortunately, it suffers from shrinking of non-zero parameters. To overcome these problems several variations of LASSO have been developed and studied, such as adaptive LASSO [34], SCAD [13] and so on. We shall now discuss a Bayesian alternative which, to some extent, provides a better tradeoff between sparsity and shrinking than the ℓ_1 norm.

This Bayesian procedure goes under the name of Sparse Bayesian Learning and can be seen as an extension of the Bayesian procedure for regularization described in the previous section. In order to illustrate the method we consider its simplest instance. Consider a discrete time MIMO FIR system as in (1) with $p = 1$ and $m = 2$; define $h_i := [h_{i1}(1),...,h_{i1}(T)]^\top$. Let $h := [h_1^\top\ h_2^\top]^\top$ and assume that the h_i's are independent Gaussian random vectors with zero mean and covariances $\eta_i K$. Letting $\Phi_i := [\phi_{1,i},..,\phi_{N,i}]^\top$ and following the formulation in (3) and (6), it follows that the marginal likelihood estimator of η takes the form

$$\begin{aligned}\hat{\eta} &:= \arg\min_{\eta_i\geq 0}\ \log(\det(\Sigma(\eta)) + y^\top\Sigma^{-1}(\eta)y \\ \Sigma(\eta) &:= \eta_1\Phi_1 K\Phi_1^\top + \eta_2\Phi_2 K\Phi_2^\top + \sigma^2 I\end{aligned} \tag{22}$$

The estimator of h is found in closed form inserting $\hat{\eta}$ in per equation (12). It can be shown that under certain conditions on the observation vector y, the estimated hyperparameters $\hat{\eta}_i$ lie at the boundary, i.e. are exactly equal to zero. If $\hat{\eta}_i = 0$ then, from equation (12), also $\hat{h}_i = 0$; this reveals that the i-th input does not enter into the model; see also Example 1 for a simple illustration.

These Bayesian methods for sparsity have been studied in a general regression framework in [32] under the name of "Type-II" Maximum Likelihood. Further results can be found in [1] which suggest that these Bayesian methods provide a better tradeoff between sparsity and shrinking (i.e. are able to provide sparse solution without inducing excessive shrinkage on the non-zero parameters).

Remark 2: A more detailed analysis, see for instance [1], shows that Lasso/GLasso (i.e. ℓ_1 penalties) and SBL using the "Empirical Bayes" approach can be derived under a common Bayesian framework starting from the joint posterior $p(\eta,h|y)$. While SBL is derived from the maximization η of the marginal posterior, Lasso/GLasso correspond to maximizing the joint posterior after a suitable change of variables. For reasons of space we refer the interested reader to the literature for details.

Recent work on the use of sparseness for variable selection and model order estimation can be found in [30], [8] and references therein.

Example 1: In order to illustrate how Sparse Bayesian Learning leads to sparse solution we consider a very simplified scenario in which the measurements equation is

$$y_t = h\,u_{t-1} + e_t$$

where e_t is zero mean, unit variance Gaussian and white and u_t is a deterministic signal. The purpose is to estimate the coefficient h, which could be possibly equal to zero. Thus the estimator should reveal whether u_{t-1} influences y_t or not.

Following the SBL framework, we model h as a Gaussian random variable, with zero mean and variance $K_\eta := \eta$,

[5]Strictly speaking any full rank FIR model of length T has Mc-Millan degree $T \times p$. Yet, we consider $\{h(k)\}_{k=1,..,T}$ to be the truncation of some "true" impulse response $\{h(k)\}_{k=1,..,\infty}$ and, as such, the finite Hankel matrix built with the coefficients $h(k)$ will have rank equal to the McMillan degree of $H(q) = \sum_{k=1}^{\infty} h(k)z^{-k}$.

[6]It is interesting to observe that both ℓ_1 and group-ℓ_1 are special cases of the nuclear norm if one considers matrices with fixed eigenspaces.

independent of e_t. Therefore y_t is also Gaussian, zero mean and variance $u_{t-1}^2\eta+1$. Therefore, assuming N data points are available, the log-likelihood function for η is given by

$$-2\log p_\eta(Y) \propto \sum_{i=1}^{N} \log(u_{t-1}^2\eta+1) + \sum_{i=1}^{N} \frac{y_t^2}{u_{t-1}^2\eta+1}$$

It is a simple to see that $\hat{\eta} := \arg\min_{\eta\geq 0} \; -2\log p_\eta(Y)$ has the form $\hat{\eta} = \max(0,\eta_*)$ where η_* is the solution of

$$\sum_{t=1}^{N} \frac{u_{t-1}^4\eta + u_{t-1}^2\left(1-y_t^2\right)}{u_{t-1}^2\eta+1} = 0$$

which unfortunately doesn't have a closed form solution. If however we assume that the input u_t is constant (without loss of generality say that $u_t = 1$), we obtain that

$$\eta_* = \frac{1}{N}\sum_{t=1}^{N} y_t^2 - 1 \quad \Rightarrow \quad \hat{\eta} = \max\left(0, \frac{1}{N}\sum_{t=1}^{N} y_t^2 - 1\right)$$

Clearly this is a threshold estimator which sets $\hat{\eta}$ to zero when the sample variance of y_t is smaller than $Var\{e_t\} = 1$. Thus the Empirical Bayes estimator of h, as per equation (12), is given by

$$\hat{h} = \frac{\hat{\eta}}{\sum_{i=1}^{N} u_{t-1}^2\hat{\eta}+1} \sum_{i=1}^{N} y_t u_{t-1}$$

which is clearly equal to zero when $K_{\hat{\eta}} = \hat{\eta} = 0$.

V. Summary and Future Directions

We have presented a bird's eye (and certainly incomplete) overview of regularization methods in System Identification. Even though regularization is quite an old topic we believe it is fair to say that the nontrivial interaction between regularization and system theoretic concepts provides a wealth of interesting and challenging problems, including kernel design, estimation of hyperparameters and their numerically efficient implementation. Much work needs to be done for multivariable (linear) systems but also non-linear system identification, which we have not been able to address in this short tutorial.

References

[1] A. Aravkin, J. Burke, A. Chiuso, and G. Pillonetto. Convex vs non-convex estimators for regression and sparse estimation: the mean squared error properties of ARD and GLasso. *Journal of Machine Learning Researc*, 2014.

[2] F. Bach, G. Lanckriet, and M. Jordan. Multiple kernel learning, conic duality, and the smo algorithm. In *Proceedings of the 21st International Conference on Machine Learning*, page 4148, 2004.

[3] M. Banbura, D. Giannone, and L. Reichlin. Large Bayesian VARs. *Journal of Applied Econometrics*, 25(1):71–92, 2010.

[4] T. Chen, M.S. Andersen, L. Ljung, A. Chiuso, and G. Pillonetto. System identification via sparse multiple kernel-based regularization using sequential convex optimization techniques. *Automatic Control, IEEE Transactions on*, page in press, 2014.

[5] T. Chen, A. Chiuso, G. Pillonetto, and L. Ljung. Rank-1 kernels for regularized system identification. In *IEEE Conference on Decision and Control*, 2013.

[6] T. Chen, H. Ohlsson, and L. Ljung. On the estimation of transfer functions, regularizations and Gaussian processes - revisited. *Automatica*, 48(8), 2012.

[7] A. Chiuso, T. Chen, L. Ljung, and G. Pillonetto. On the design of multiple kernels for nonparametric linear system identification. In *submitted to IEEE CDC 2014*, 2014. *http://automatica.dei.unipd.it/people/chiuso/publications.html*.

[8] A. Chiuso and G. Pillonetto. A bayesian approach to sparse dynamic network identification. *Automatica*, 48(8):1553 – 1565, 2012.

[9] G. De Nicolao, G. Ferrari Trecate, and A. Lecchini. MaxEnt priors for stochastic filtering problems. In *Proc. Symp. on Math. Theory of Networks and Systems (MTNS'98)*, pages 755–758, Padova, Italy, 1998.

[10] F. Dinuzzo. Kernels for linear time invariant system identification. *http://arxiv.org/abs/1203.4930*, 2012.

[11] T. Doan, R. Litterman, and C.A. Sims. Forecasting and conditional projection using realistic prior distributions. *Econometric Reviews*, 3:1–100, 1984.

[12] D. Donoho. Compressed sensing. *IEEE Trans. on Information Theory*, 52(4):1289–1306, 2006.

[13] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, december 2001.

[14] M. Fazel, H. Hindi, and S.P. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, 2001. Proceedings of the 2001*, volume 6, pages 4734 –4739 vol.6, 2001.

[15] H. Hjalmarsson, J. Welsh, and C.R. Rojas. Identification of box-jenkins models using structured ARX models and nuclear norm relaxation. In *16th IFAC Symposium on System Identification*, pages 322–327. IFAC, 2012.

[16] R. R. Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, 32(1):pp. 1–49, 1976.

[17] G. Kitagawa and H. Gersh. A smothness priors-state space modeling of time series with trends and seasonalities. *Journal of the American Statistical Association*, 79(386):378–389, 1984.

[18] H. Leeb and B. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21:2159, 2005.

[19] L. Ljung. *System Identification, Theory for the User*. Prentice Hall, 1997.

[20] D.J.C. Mackay. Bayesian non-linear modelling for the prediction competition. *ASHRAE Trans.*, 100(2):3704–3716, 1994.

[21] G. Pillonetto, A. Chiuso, and G. De Nicolao. Prediction error identification of linear systems: A nonparametric Gaussian regression approach. *Automatica*, 47:291–305, 2011.

[22] G. Pillonetto, A. Chiuso, and G. De Nicolao. Convexifying missing data problems using maximum entropy kernels for linear system identification. Technical report, University of Padova, 2014. [ONLINE] Submitted to IEEE Trans. on Aut. Control.

[23] G. Pillonetto and G. De Nicolao. A new kernel-based approach for linear system identification. *Automatica*, 46:81–93, 2010.

[24] G. Pillonetto, F. Dinuzzo, T. Chen, G. De Nicolao, and L. Ljung. Kernel methods in system identification, machine learning and function estimation: a survey. *Automatica*, 50:657–682, 2014.

[25] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[26] T. Söderström and P. Stoica. *System Identification*. Prentice-Hall, 1989.

[27] R. Tibshirani. Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society, Series B.*, 58:267–288, 1996.

[28] M. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

[29] G. Wahba. *Spline models for observational data*. SIAM, Philadelphia, 1990.

[30] H. Wang, G. Li, and C.L. Tsai. Regression coefficient and autoregressive order shrinkage and selection via the lasso. *Journal Of The Royal Statistical Society Series B*, 69(1):63–78, 2007.

[31] N. Wiener. *The Fourier Integral and Certain of its Applications*. Cambridge Univ. Press, 1933.

[32] D.P. Wipf, B.D. Rao, and S. Nagarajan. Latent variable Bayesian models for promoting sparsity. *IEEE Transactions on Information Theory*, 57(98):6236 – 6255, 2011.

[33] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

[34] H. Zou. The adaptive Lasso and it oracle properties. *Jorunal of the American Statistical Association*, 101(476):1418–1429, 2006.